STAT347: Generalized Linear Models
Lecture 1

Today's topics: Agresti Chapter 1

- Two real data examples
- GLM concepts

1 Two real data examples

Please check the R notebook 1.

2 Components of a GLM

Data points $(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_n)$

1. Random components:
   Observations $(y_1, y_2, \cdots, y_n)$ follow some distribution family and are independent
   - Generalize y_i from continuous real values to binary response, counts, categories, et. al.
   - How to describe the distribution of y?
     We will start with assuming y_i coming from an exponential family distribution.
   - Treat the covariates $(X_1, \cdots, X_n)$ as fixed. For random X, build the model conditional on X.
2. Link function:
   $g(\mathbb{E}(y_i)) = g(\mu_i) = X_i^T\beta$ where $\beta = (\beta_1, \cdots, \beta_p)^T$ and $X_i = (x_{i1}, \cdots, x_{ip})^T$
   - linear model: $g(\mu_i) = \mu_i$
   - model for counts: $g(\mu_i) = \log(\mu_i)$.
   - model for binary data: $g(\mu_i) = g(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$.
3. linear predictor:
   $X\beta$ where $X = (X_1, X_2, \cdots, X_n)^T$ is the $n \times p$ model matrix.
   - X can include interactions, non-linear transformations of the observed covariates and the constant term
   - avoid causal interpretations of the coefficients β (read Chapter 1.2.3)

3 GLM v.s. data transformation

An alternative to GLM is to transform y_i in some $h(y_i)$ and build a linear model $h(y_i) = X_i^T \beta + \epsilon_i$.

- Sounds a reasonable approach, and is still commonly used now in various applications.
- If y_i are counts, usually take $h(y_i) = \log(y_i)$. How to deal with $y_i = 0$? How to transform binary or categorical data? Also, the variance is not stabilized after transformation.
- Disadvantage of data transformation: need to find h that can make a linear model reasonable as well as stabilizing the variance. (read Chapter 1.1.6)
- Advantage of data transformation in practice: easier to build models more complicated than a regression model if we think the transformed data are approximately Gaussian.

Next time: Agresti Chapters 4.1-4.2, exponential family distribution, ML estimation of GLM